

الذخائر اللغوية ومعالجة النصوص التراثية: تحديات وآفاق في عصر الذكاء الاصطناعي

Linguistic Corpora and Heritage Text Processing: Challenges and Prospects in the Era of Artificial Intelligence

سمير جاب الله*

جامعة محمد بوضياف- المسيلة- الجزائر

samir.djaballah@univ-msila.dz

ملخص: (لا يتجاوز 10 أسطر)	معلومات المقال
تتناول هذه الدراسة التحليلية دور الذخائر اللغوية (Corpora) بوصفها بنية تحتية لا غنى عنها في معالجة اللغة العربية آلياً، مع التركيز على النصوص التراثية التي تجسد الذاكرة الحضارية للأمة وتعكس عمق هويتها. ويسعى البحث إلى رصد جملة من التحديات الجوهرية التي تعترض رقمنة المخطوطات ومعالجتها، من بينها: غياب التشكيل، وتعدد النسخ الخطية المتباينة، والتعقيد الصرفي والدلالي الكامن في أبنية العربية. كما يقرئ المقال القفزات النوعية التي أحرزتها تقنيات الذكاء الاصطناعي والتعلم العميق في السنوات الأخيرة، مستشهداً بمشاريع ريادية ومنصات وطنية، في مقدمتها منصتا (فلك) و(سوار) التابعتان لمجمع الملك سلمان العالمي للغة العربية. وينتهي البحث برسم آفاق مستقبلية لتوظيف الذكاء الاصطناعي التوليدي في تحقيق النصوص التراثية وترميمها رقمياً.	تاريخ الارسال:
	2026-03-31
	تاريخ القبول:
	2026-06-12
	الكلمات المفتاحية: ✓ الذخائر اللغوية ✓ الكلمات النصوص التراثية ✓ الذكاء الاصطناعي
Abstract : (not more than 10 Lines)	Article info

This analytical study examines the pivotal role of linguistic corpora as an indispensable infrastructure for Arabic Natural Language Processing (NLP), focusing on heritage texts that embody cultural memory and national identity. The research identifies core challenges in manuscript digitization, including the absence of diacritics, divergent manuscript versions, and the inherent morphological and semantic complexities of the Arabic language. Furthermore, it explores qualitative leaps in AI and Deep Learning, citing pioneering national platforms such as "Falak" and "Siwar" by the King Salman Global Academy for Arabic Language. The study concludes by outlining future prospects for employing Generative AI in the digital verification and restoration of heritage materials, ensuring their preservation within a modern technological framework.

Keywords:

- ✓ LinguisticCorpora
- ✓ HeritageTexts
- ✓ Artificial Intelligence (AI)

مقدمة

شهدت اللسانيات الحاسوبية في العقد الأخير تحولاً بنوياً عميقاً؛ إذ انتقل هذا الحقل المعرفي من الاشتغال بالقواعد الصورية الضيقة إلى رحاب النمذجة الإحصائية والشبكات العصبية العميقة. وفي قلب هذا التحول تتربع "الذخيرة اللغوية (Corpus)" مكانةً محوريةً، لا بوصفها مستودعاً نصياً سلبياً فحسب، بل كياناً حياً يغذي خوارزميات الذكاء الاصطناعي بما تحتاج إليه من بيانات تدريبية دقيقة تعكس حقيقة اللغة في تنوعها وثرائها.

وتكتسي معالجة النصوص التراثية العربية ألياً أهميةً مزدوجة الأبعاد؛ فهي من جهة تمثل تحدياً تقنياً خالصاً نظراً لما تنطوي عليه اللغة العربية من خصائص بنوية وتاريخية بالغة التعقيد، ومن جهة ثانية تُعدّ ضرورةً حضاريةً لا تقبل التأجيل؛ إذ يغدو إتاحة هذا الكنز المعرفي للأجيال القادمة وللباحثين في شتى أنحاء العالم التزاماً أخلاقياً تجاه الموروث الإنساني المشترك. ومع توسّع نماذج اللغة الضخمة (LLMs) وانتشارها في الفترة الأخيرة، بات من الملح إعادة تقييم منهجيات بناء الذخائر التراثية لضمان دقة الاسترجاع وعمق التحليل اللساني وصحة التمثيل.

1. مفهوم الذخائر اللغوية وأهميتها في الدراسات التراثية

وسع الإطار النظري والمنهجي لهذا الحقل البحثي ليشمل أبعاداً تحليلية أكثر شمولاً في دراسة النصوص التراثية، من خلال الانتقال من التصور الخطي للمدونة اللغوية إلى تصور شبكي متعدد الطبقات، يُنظر فيه إلى النص بوصفه بنية ديناميكية تتقاطع فيها مستويات اللغة مع سياقات الإنتاج والتلقي وإعادة التوظيف عبر الزمن. وفي هذا السياق، يغدو التحليل اللغوي جزءاً من منظومة أوسع

تُعنى بـ"الذاكرة النصية" التي لا تختزل في المفردات أو التراكيب، بل تمتد إلى أنماط التفكير وأنساق المعرفة التي أنتجت تلك النصوص داخل بيئاتها التاريخية.

كما يقتضي هذا التوسيع إعادة بناء الإطار المنهجي على أساس التكامل بين التحليل الكمي والتحليل الكيفي؛ إذ لا يكفي الاكتفاء بالإحصاء التكراري أو النمذجة الاحتمالية، بل ينبغي ربط النتائج الكمية بتأويلات دلالية وسياقية تستند إلى علوم التراث ذاتها، مثل علم البلاغة وأصول الفقه والنقد الأدبي. ومن ثمّ يصبح تحليل النصوص التراثية عملية مزدوجة تقوم على تفكيك البنى اللغوية من جهة، وإعادة تركيبها داخل سياقاتها المعرفية الأصلية من جهة ثانية، بما يسمح باستعادة المعنى في تعدديته لا في صورته المبسطة (الشحات، 2022، ص. 222).

ال وفي مستوى آخر، يفرض هذا الإطار الموسّع إدماج تقنيات الذكاء الاصطناعي التوليدي والنماذج اللغوية الضخمة ضمن سلسلة التحليل، ليس بوصفها أدوات تقنية فقط، بل بوصفها فاعلاً منهجياً يعيد تشكيل طريقة التعامل مع النصوص. غير أن هذا الإدماج يستوجب وضع ضوابط علمية صارمة لضمان عدم طمس الخصوصية التراثية أو اختزالها في تمثيلات إحصائية مجردة، وذلك عبر اعتماد نماذج هجينة تجمع بين المعرفة القاعدية المستمدة من المصادر التراثية والخوارزميات التكيفية الحديثة.

ومن منظور أبعد، يمكن توسيع مفهوم "التمثيلية" ليشمل تمثيلية معرفية وثقافية، بحيث لا تقتصر الذخائر اللغوية على جمع النصوص وفق معايير زمنية أو كمية، بل تتجه نحو بناء مدونات تعكس تنوع المدارس الفكرية، وتعدد البيئات العلمية، واختلاف أنماط التداول المعرفي داخل الحضارة العربية الإسلامية. وهذا ما يسمح بإعادة قراءة التراث بوصفه شبكة تفاعلية من المعارف المتداخلة، لا مجرد أرشيف لغوي ثابت (بنزينة ودريم، 2019، ص. 22).

وبذلك يتحول هذا الحقل البحثي إلى فضاء معرفي مركب، تتداخل فيه اللسانيات الحاسوبية مع علوم التاريخ والفكر، وتندمج فيه الخوارزميات مع المناهج التأويلية، ليصبح الهدف النهائي ليس فقط معالجة النصوص التراثية، بل إعادة بناء فهم شامل لها بوصفها منظومة معرفية حية قابلة للتحليل وإعادة الإنتاج في السياق الرقمي المعاصر (الحناش، 2015).

تُعدّ المعالجة الآلية للنصوص التراثية العربية من أشدّ الملفات تعقيداً في اللسانيات الحاسوبية المعاصرة؛ إذ تتشابك تعقيدات الحامل المادي للنص مع عقبات لغوية بنيوية متجذّرة.

وتتصدر هذه التحديات مسألة "التعرف الضوئي على الحروف (OCR) و"التعرف على الكتابة اليدوية (HTR)"، إذ تعاني المخطوطات القديمة من تآكل الأوراق وتداخل الأسطر وانعدام المسافات المنتظمة بين الكلمات، إضافة إلى تنوع الأنماط الخطية من الكوفي إلى النسخ والتلث، وهي أنماط تفتقر في الغالب إلى ما يُعرف بـ"الخط القاعدي الثابت (Consistent Baseline)" مجلة الإيسيسكو للغة العربية، 2024، ص. 272).

2.1 غياب التشكيل واللبس الدلالي

يمثل "التشكيل الآلي (Automatic Diacritization)" تحدياً لسانياً معقداً؛ فغياب الحركات الإعرابية يجعل الكلمة الواحدة قابلة لعدة قراءات دلالية. وقد أشار الحاج صالح إلى أن التشكيل ليس مسألة تقنية فقط، بل إشكالية معرفية (الحاج صالح، 2007).

2.2 التطور الدلالي والمصطلحي

تُعدّ ظاهرة تغير المعاني التاريخي من أصعب التحديات، إذ قد تحمل الكلمة في النص التراثي دلالة تختلف جذرياً عن استعمالها الحديث (الشحات، 2022، ص. 227).

2.3 تعدد النسخ الخطية

إن تعدد النسخ الخطية للنص الواحد يزيد من درجة التعقيد النصي، ويستدعي تطوير نماذج حاسوبية متقدمة قادرة على تمثيل هذا التعدد في صورة بنى طبقية، تسمح بالمقارنة الآلية بين الشواهد المختلفة، واستخراج الفروقات الدقيقة في المستوى اللفظي والدلالي دون الإخلال بالبنية المعنوية الأصلية للنص (بنزينة ودريم، 2019، ص. 25). ويتطلب ذلك الانتقال من مفهوم "النص الواحد الثابت" إلى مفهوم "النص المتغير تاريخياً"، أي التعامل مع المخطوط بوصفه كياناً حياً متعدد الصيغ، لا وثيقة جامدة، مما يفتح المجال أمام نماذج التحليل المقارن للمخطوطات (Textual Variants Analysis) التي تدمج بين الرؤية الحاسوبية والمعايير النقدية التراثية في التحقيق.

3. الذكاء الاصطناعي والتعلم العميق: ثورة في معالجة التراث

أفضى توظيف نماذج المحولات (Transformers) وآليات الانتباه (Attention Mechanisms) إلى قفزة نوعية في فهم السياق النصي الطويل، حيث لم يعد النموذج يعتمد على المجاورة المحلية للكلمات، بل أصبح قادرًا على تتبع العلاقات الدلالية عبر مسافات نصية ممتدة، وهو ما أتاح فهمًا أكثر دقة للبنى الحجاجية والأسلوبية في النصوص الكلاسيكية. (Chan, 2024, p. 1) وقد أسهم هذا التطور في إعادة تعريف مفهوم "الفهم الآلي للنص"، من مجرد معالجة إحصائية إلى تمثيل سياقي عميق للبنية المعرفية للنص.

وفي السياق ذاته، أتاح نموذج AraModernBERT القدرة على استيعاب سياقات أطول ومعالجة تراكيب لغوية معقدة، مما انعكس إيجابًا على تحليل النصوص التراثية التي تتميز بتداخل مستوياتها النحوية والدلالية، مثل النصوص الفقهية والتفسيرية التي تعتمد على ترابط استدلالى ممتد. (Elshehy et al., 2026, p. 2). كما ساهم نظام "سكون" في معالجة الغموض المعجمي واستعادة التشكيل، عبر نماذج احتمالية تأخذ بعين الاعتبار السياق التاريخي واللغوي للكلمة، مما يقلل من أخطاء التأويل الناتجة عن غياب الحركات في المخطوطات. (KSAA, 2026)

وبشكل عام، لم يعد الذكاء الاصطناعي مجرد أداة تقنية مساعدة، بل أصبح فاعلاً تحليليًا يعيد تشكيل طريقة قراءة التراث، من خلال الكشف عن الأنماط الخفية، وربط النصوص ببعضها ضمن شبكات دلالية واسعة، تسمح بإعادة بناء المعرفة التراثية بطريقة أكثر تكاملاً وعمقاً.

4. مبادرات ريادية

تشكل منصات مجمع الملك سلمان العالمي للغة العربية، مثل "فلك" و"سوار"، بنية تحتية معرفية متقدمة تهدف إلى تحويل اللغة العربية من مادة نصية مشتتة إلى موارد رقمية منظمة قابلة للاستثمار في نماذج الذكاء الاصطناعي. (KSAA, 2024–2026) فهذه المنصات لا تقتصر على توفير البيانات، بل تقدم بيانات تحليل متكاملة تشمل أدوات الإحصاء اللغوي، والتحليل السياقي، وبناء المعاجم الديناميكية، مما يجعلها جزءاً من منظومة أوسع لإعادة هندسة المعرفة اللغوية العربية.

أما مشروع "زنكي"، فيمثل خطوة متقدمة في مجال رقمنة الشبكات المعرفية التاريخية، حيث يعتمد على تقنيات استخراج الكيانات المسماة (NER) وربطها بالعلاقات الزمنية والجغرافية والعلمية، بما يسمح بإعادة بناء خرائط المعرفة الإسلامية عبر العصور. ومن خلال هذا النهج، لا يُنظر إلى التراث بوصفه نصوصًا منفصلة، بل بوصفه شبكة تفاعلية من العلماء والأفكار والمؤسسات العلمية (ZinkiResearch Group, 2024).

5. آفاق المستقبل

يتجه المستقبل نحو تطوير ما يمكن تسميته بـ"المحقق الرقمي الآلي"، وهو نظام هجين يجمع بين الذكاء الاصطناعي والتحقيق النقدي التقليدي، بحيث لا يكتفي بمقارنة النسخ الخطية، بل يقترح ترميمات نصية مدعومة بأدلة احتمالية مستمدة من الاستخدامات اللغوية التاريخية وأنماط التأليف عند المؤلفين (الزبيدي، 2025، ص. 232). وهذا التحول من "المعالجة" إلى "التحقيق الذكي" يمثل نقلة نوعية في الدراسات التراثية.

كما يُتوقع أن يشهد المستقبل دمجًا أعمق بين تقنيات الواقع المعزز (AR) والواقع الافتراضي (VR) وبين الذخائر اللغوية، بحيث يتمكن الباحث من التفاعل مع المخطوطات في بيئات ثلاثية الأبعاد، مع طبقات تفسيرية فورية تشمل التشكيل، والشرح الدلالي، والسياق التاريخي. إضافة إلى ذلك، فإن تفعيل التشغيل البيئي الدلالي عبر البيانات المترابطة (Linked Data) سيتيح ربط النصوص التراثية بقواعد بيانات عالمية، مما يحول التراث العربي إلى جزء فاعل في الشبكة المعرفية الإنسانية الكبرى (جامعة السلطان قابوس، 2024، ص. 19).

وبذلك، يتضح أن مستقبل معالجة التراث لم يعد يقتصر على الرقمنة أو الفهرسة، بل يتجه نحو إعادة بناء المعرفة التراثية نفسها بوصفها نظامًا ديناميكيًا قابلاً للتحليل، وإعادة التفسير، وإعادة الإنتاج داخل الفضاء الرقمي العالمي.

خاتمة

إن بناء الذخائر اللغوية للنصوص التراثية يمثل ضرورة استراتيجية ملحة لصون الهوية اللغوية والثقافية في العصر الرقمي، وفي الوقت ذاته يشكل مدخلًا أساسيًا لإعادة إدماج التراث العربي في

منظومة إنتاج المعرفة المعاصرة. فهذه الذخائر لم تعد مجرد أدوات تقنية لحفظ النصوص أو معالجتها، بل تحولت إلى بنى معرفية قادرة على إعادة تشكيل علاقتنا بالموروث، عبر تحويله من مادة محفوظة في الرفوف والأرشفات إلى بيانات ديناميكية قابلة للتحليل والاستثمار العلمي.

ولا يتحقق هذا التحول النوعي إلا من خلال تكامل الجهود بين اللسانيين، بما يمتلكونه من أدوات تحليل لغوي وتاريخي، والمهندسين وخبراء الذكاء الاصطناعي، بما يطورونه من نماذج وخوارزميات قادرة على التعامل مع التعقيد البنيوي للنصوص التراثية. كما أن هذا التكامل ينبغي أن يكون مؤطرًا برؤية علمية مشتركة تراعي خصوصية اللغة العربية وتعدد مستوياتها التاريخية والأسلوبية، وتضمن في الوقت نفسه دقة المعالجة الحاسوبية وموثوقية النتائج.

وفي هذا السياق، تبرز أهمية المشاريع المؤسسية الرائدة مثل منصتي "فلك" و"سوار"، بوصفها نماذج عملية لتجسيد هذا التكامل بين المعرفة اللغوية والتقنيات الحديثة، حيث تساهم في بناء قواعد بيانات ضخمة، وتوفير أدوات تحليل متقدمة، وتيسير الوصول إلى المعاجم والنصوص التراثية بصورة معيارية ومنظمة. إن هذه المبادرات لا تمثل مجرد جهود تقنية معزولة، بل تشكل مكونات ضمن مشروع حضاري أوسع يهدف إلى إعادة تموضع اللغة العربية في قلب التحولات الرقمية العالمية.

وبذلك، يغدو التراث العربي من خلال الذخائر اللغوية موردًا معرفيًا حيًا، لا يقتصر دوره على حفظ الذاكرة الثقافية للأمة، بل يمتد ليصبح عنصرًا فاعلاً في إنتاج المعرفة الحديثة، وإسهامًا مباشرًا في بناء مستقبل رقمي تتداخل فيه الهوية مع التكنولوجيا، والأصالة مع الابتكار.

الهوامش

• الحاج صالح، عبد الرحمن. (2007). مشروع الذخيرة اللغوية العربية وأبعاده العلمية والتطبيقية. مجلة اللغة العربية، مج. 3، ع. 1، ص. 9-30.

• الشحات، أسماء. (2022). المدونات اللغوية ودورها في المعجمية العربية الحديثة. دار الوفاء.

• شريف، كامل شاهين. (2024). الدليل العربي لتكامل مؤسسات ذاكرة المجتمعات: المكتبات والأرشفات والمتاحف ومتطلبات التحول الرقمي. جامعة السلطان قابوس، مسقط.

• مبادرة الملك عبد الله للمحتوى العربي. (2025). تقرير الذكاء الاصطناعي والبيانات اللغوية العربية: نحو مليار كلمة لتدريب النماذج الكبيرة. الرياض، ص. 15.

• مجمع الملك سلمان العالمي للغة العربية. (2026). (a) السيادة الرقمية وواجهات برمجة المعاجم التراثية: منصة سوار نموذجًا. مركز ذكاء العربية، الرياض، ص. 10.

المقالات العلمية:

• بنزينة، صفية؛ ودریم، نور الدين. (2019). مشروع الذخيرة اللغوية العربية: الواقع والآفاق. مجلة لغويات، 4(1)، ص. 15-30.

• الحناش، محمد. (2015). استخدام اللغة العربية في تقنية المعلومات. مجلة التواصل اللساني، فاس.

• الزيدي، و. ك. (2025). استخدامات الذكاء الاصطناعي في فك شفرات الكتابات القديمة وترميم المخطوطات التاريخية المتضررة. أعمال المؤتمر العلمي الدولي السادس لتاريخ العلوم عند العرب، ص. 220-235.

• العمران، فاطمة مصطفى. (2026). الشعر الجاهلي بين الشروح التراثية والتحليل الآلي: دراسة لغوية باستخدام تقنيات التعلم العميق. مجلة بنك، 6(1)، ص. 1-20.

• مجلة الإيسيسكو للغة العربية. (2024). آفاق الذكاء الاصطناعي في خدمة اللغة العربية: المخطوطات أنموذجًا. منظمة العالم الإسلامي للتربية والعلوم والثقافة، العدد 12، الرباط.

- Alayba, A. M. (2025). Arabic natural language processing (NLP): A comprehensive review of challenges, techniques, and emerging trends. *Computers*, 14(11), 497.
- Chan, A., Mijar, A., Saeed, M., Wong, C.-W., & Khater, A. (2024). HATFormer: Historic handwritten Arabic text recognition with transformers. *arXiv preprint arXiv:2410.02179*, 1-15.
- Elshehy, O., Nacar, O., Djamai, A., Ragab, M., Al Jallad, K., et al. (2026). AraModernBERT: Transtokenized initialization and long-context encoder modeling for Arabic. *arXiv preprint arXiv:2603.09982*, 1-12.
- Zinki Research Group. (2024). Mapping the Islamic Republic of Letters: Named entity extraction and relational networks in classical Arabic texts. *Working Paper Series*, p. 8.

• مواقع الانترنت:

- مجمع الملك سلمان العالمي للغة العربية. (2024). منصة سوار للمعاجم اللغوية. الرياض. مسترجع من: <https://siwar.ksaa.gov.sa>
- مجمع الملك سلمان العالمي للغة العربية. (2024). منصة فلك للمدونات اللغوية. الرياض. مسترجع من: <https://falak.ksaa.gov.sa>
- مجمع الملك سلمان العالمي للغة العربية. (2026). تقرير المهمة المشتركة KSAA-2026 للتشكيل الآلي. مركز ذكاء العربية، الرياض <https://arai.ksaa.gov.sa/arsharedTask2026/>
- مكتبة قطر الوطنية. (2024). تقرير رقمنة المخطوطات العربية وتحديات التعرف الآلي. الدوحة. <https://www.qnl.qa/ar/about/annual-report>

- مؤتمر حوسبة اللغة العربية وإثراء البيانات اللغوية. (2024). أعمال المؤتمر الدولي الأول للمجمع الرياض.